Seeing with Algorithms: Deep Dive into Object Detection

From Classification to Localization and Detection Metrics

Nipun Batra and teaching staff

IIT Gandhinagar

August 5, 2025

Learning Outcomes

- Understand **classification**, **localization**, and **detection**

"Detection is not just about finding objects, but finding them right."

Learning Outcomes

- Understand **classification**, **localization**, and **detection**
- Master **precision**, **recall**, **AP**, **mAP**, and **CA-mAP**

"Detection is not just about finding objects, but finding them right."

Learning Outcomes

- Understand **classification**, **localization**, and **detection**
- Master **precision**, **recall**, **AP**, **mAP**, and **CA-mAP**
- Build strong intuition with toy examples and visual explanations

"Detection is not just about finding objects, but finding them right."

Learning Outcomes

- Understand **classification**, **localization**, and **detection**
- Master **precision**, **recall**, **AP**, **mAP**, and **CA-mAP**
- Build strong intuition with toy examples and visual explanations
- Learn to evaluate object detectors thoroughly and effectively

"Detection is not just about finding objects, but finding them right."

Roadmap

1. Motivation and Applications
2. What is Object Detection?
3. Our 3-Class Detection Example
4. Detection Pipeline
5. Evaluation Metrics: The Foundation
6. Precision-Recall Curves and Average Precision
7. Mean Average Precision (mAP)
8. Advanced Topics

Task Definitions

Definition: Three Fundamental Computer Vision Tasks

- **Classification**: What is present in the image?

Each task builds upon the previous one, increasing in complexity and practical utility.

Task Definitions

Definition: Three Fundamental Computer Vision Tasks

- **Classification**: What is present in the image?
- **Localization**: Where is the object in the image?

Each task builds upon the previous one, increasing in complexity and practical utility.

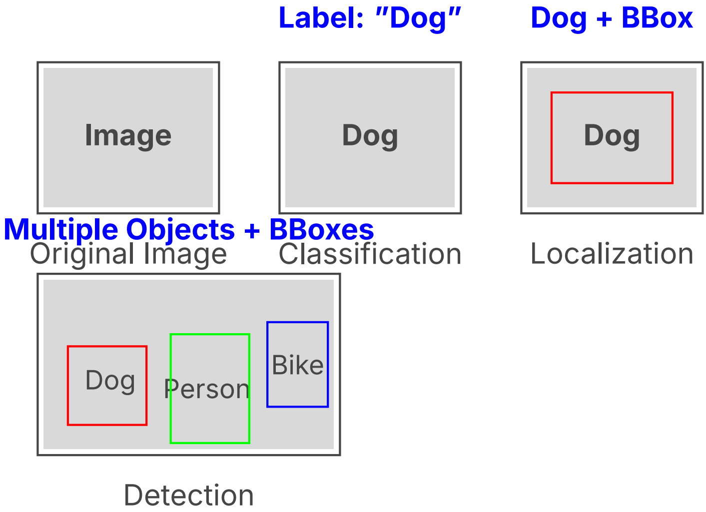
Task Definitions

Definition: Three Fundamental Computer Vision Tasks

- **Classification**: What is present in the image?
- **Localization**: Where is the object in the image?
- **Detection** = Classification + Localization (for multiple objects)

Each task builds upon the previous one, increasing in complexity and practical utility.

Visual Examples



Output Formats

Task	Output Format	Example
Classification	`label`	`"dog"`
Localization	`label, bbox`	`"dog", (30,3`

Detection	`[label, conf, bbox] × N`	`["dog", 0.95`
		`["person", 0`
		`["bike", 0.7`

Key Points

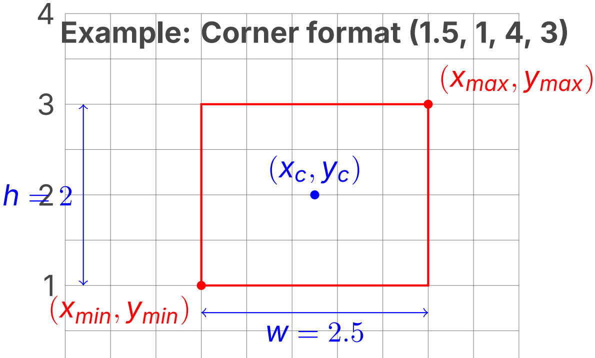
Key Insight: Detection outputs include confidence scores, enabling ranking and threshold-based filtering!

What is a Bounding Box?

Definition: Bounding Box Formats

- **Corner format:** $(x_{min}, y_{min}, x_{max}, y_{max})$

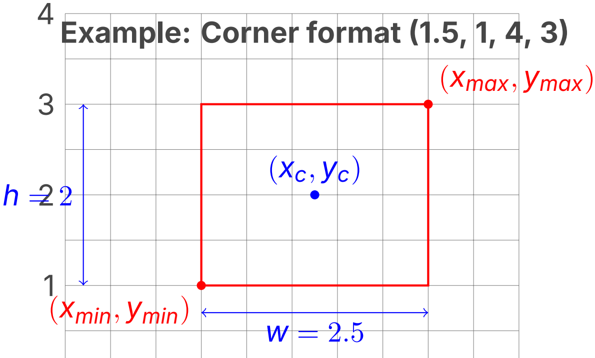
Example: Corner format (1.5, 1, 4, 3)



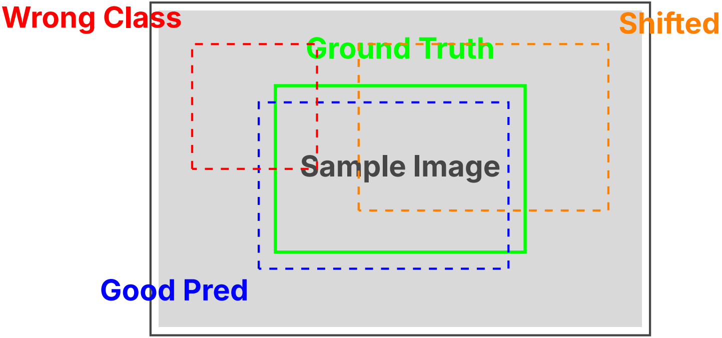
What is a Bounding Box?

Definition: Bounding Box Formats

- **Corner format:** $(x_{min}, y_{min}, x_{max}, y_{max})$
- **Center format:** $(x_{center}, y_{center}, width, height)$



Ground Truth vs Predictions



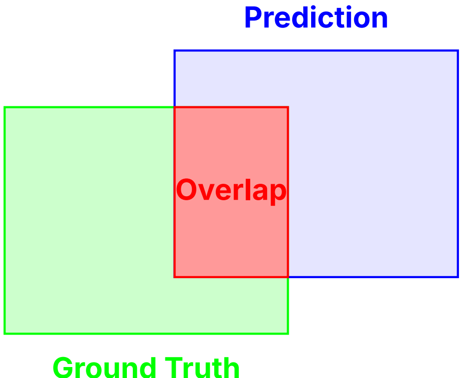
Key Points

Matching Question: How do we decide which predictions correspond to which ground truth objects?

IoU Definition

Definition: Intersection over Union (IoU)

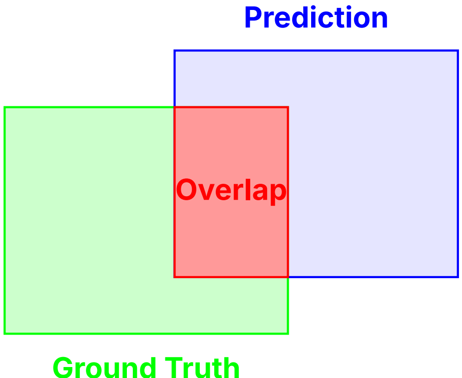
$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|A \cap B|}{|A \cup B|}$$



IoU Definition

Definition: Intersection over Union (IoU)

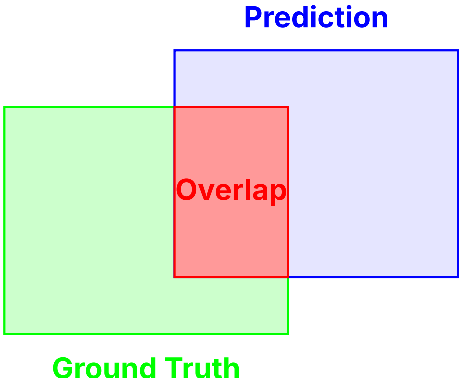
$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|A \cap B|}{|A \cup B|}$$



IoU Definition

Definition: Intersection over Union (IoU)

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|A \cap B|}{|A \cup B|}$$



IoU Calculation Example

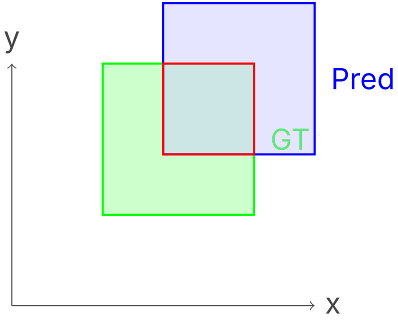
Example: Step-by-Step IoU Calculation

Ground Truth: $(30, 30, 100, 100)$
Prediction: $(50, 50, 120, 120)$

IoU Calculation Example

Example: Step-by-Step IoU Calculation

Ground Truth: $(30, 30, 100, 100)$
Prediction: $(50, 50, 120, 120)$

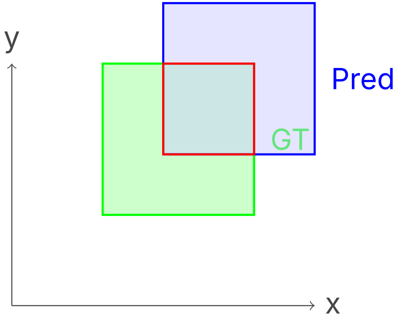


Scale: 1 unit = 20 pixels

IoU Calculation Example

Example: Step-by-Step IoU Calculation

Ground Truth: $(30, 30, 100, 100)$
Prediction: $(50, 50, 120, 120)$



Scale: 1 unit = 20 pixels

Step 1: Find intersection

$x_{min} = \max(30, 50) = 50$
$y_{min} = \max(30, 50) = 50$
$x_{max} = \min(100, 120) = 100$
$y_{max} = \min(100, 120) = 100$

Step 2: Calculate areas

Intersection: $50 \times 50 = 2500$
GT area: $70 \times 70 = 4900$
Pred area: $70 \times 70 = 4900$
Union:
$4900 + 4900 - 2500 = 7300$

Definitions

Definition: Core Metrics

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

- **Precision**: What fraction of detections are correct? (Quality)

Key Points

Definitions

Definition: Core Metrics

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

- **Precision**: What fraction of detections are correct? (Quality)
- **Recall**: What fraction of ground truth objects are detected? (Coverage)

Key Points

Definitions

Definition: Core Metrics

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

- **Precision**: What fraction of detections are correct? (Quality)
- **Recall**: What fraction of ground truth objects are detected? (Coverage)
- **TP**: True Positive (correct detection, IoU $\geq$ threshold)

Key Points

Definitions

Definition: Core Metrics

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

- **Precision**: What fraction of detections are correct? (Quality)
- **Recall**: What fraction of ground truth objects are detected? (Coverage)
- **TP**: True Positive (correct detection, IoU $\geq$ threshold)
- **FP**: False Positive (incorrect detection, IoU $<$ threshold or extra detection)

Key Points

Definitions

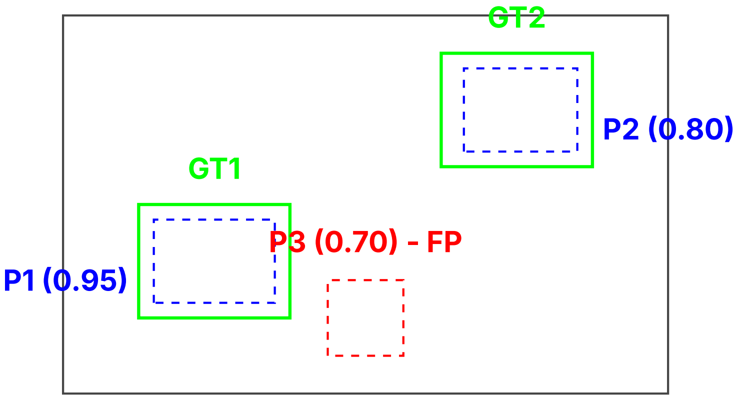
Definition: Core Metrics

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

- **Precision**: What fraction of detections are correct? (Quality)
- **Recall**: What fraction of ground truth objects are detected? (Coverage)
- **TP**: True Positive (correct detection, IoU $\geq$ threshold)
- **FP**: False Positive (incorrect detection, IoU $<$ threshold or extra detection)
- **FN**: False Negative (missed ground truth object)

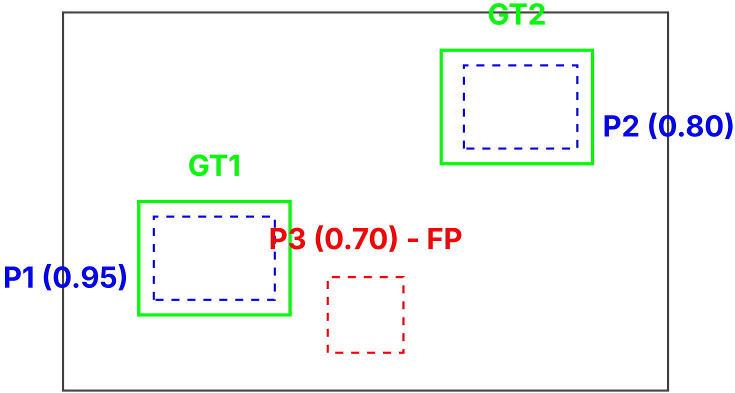
Key Points

Example: Counting TP, FP, FN



Scenario: 2 GT objects, 3 predictions

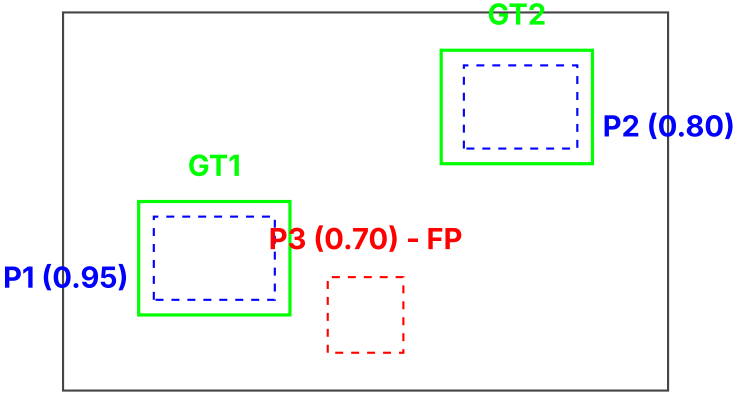
Example: Counting TP, FP, FN



Scenario: 2 GT objects, 3 predictions

Analysis (IoU threshold = 0.5):

Example: Counting TP, FP, FN

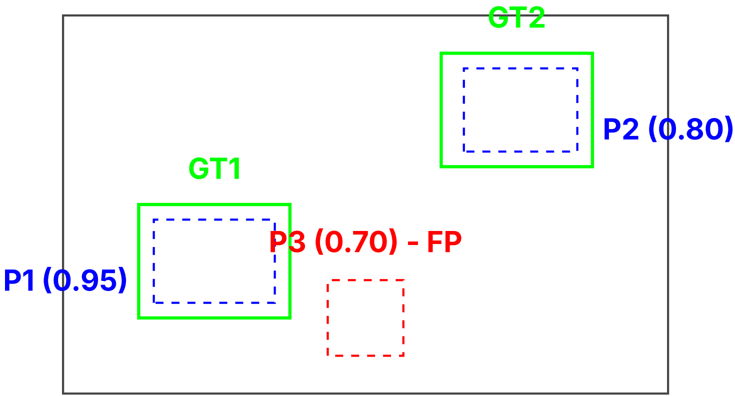


Scenario: 2 GT objects, 3 predictions

Analysis (IoU threshold = 0.5):

- P1 matches GT1: **TP**

Example: Counting TP, FP, FN

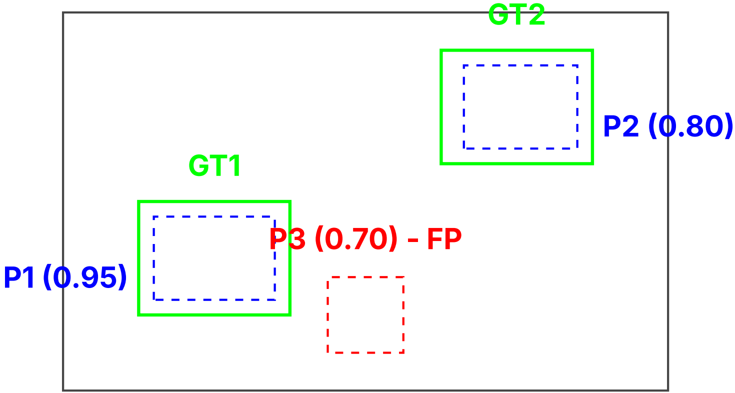


Scenario: 2 GT objects, 3 predictions

Analysis (IoU threshold = 0.5):

- P1 matches GT1: **TP**

Example: Counting TP, FP, FN

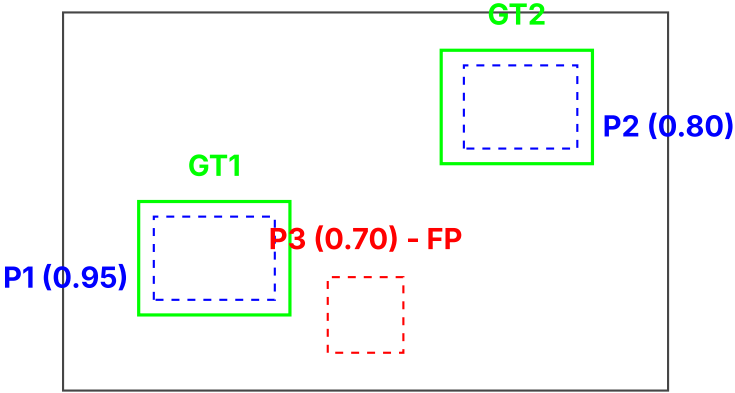


Scenario: 2 GT objects, 3 predictions

Analysis (IoU threshold = 0.5):

- P1 matches GT1: **TP**

Example: Counting TP, FP, FN

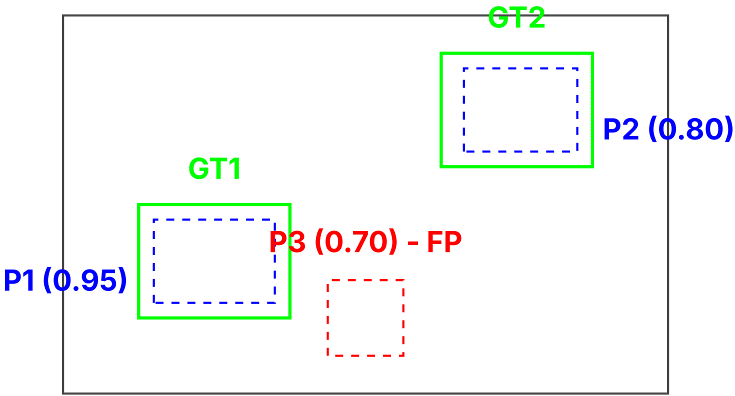


Scenario: 2 GT objects, 3 predictions

Analysis (IoU threshold = 0.5):

- P1 matches GT1: **TP**

Example: Counting TP, FP, FN



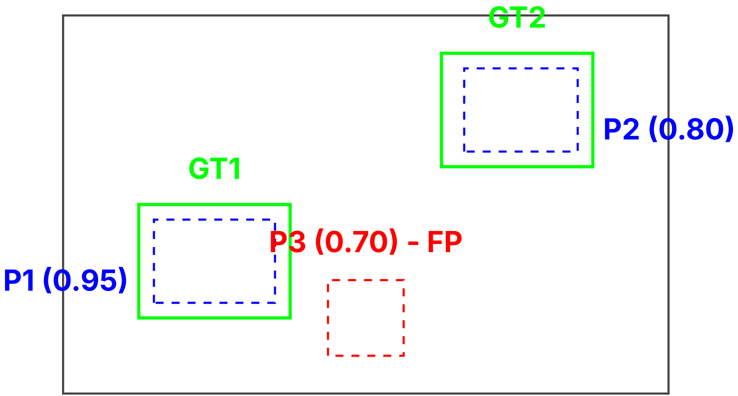
Scenario: 2 GT objects, 3 predictions

Analysis (IoU threshold = 0.5):

- P1 matches GT1: **TP**

Metrics:

Example: Counting TP, FP, FN



Scenario: 2 GT objects, 3 predictions

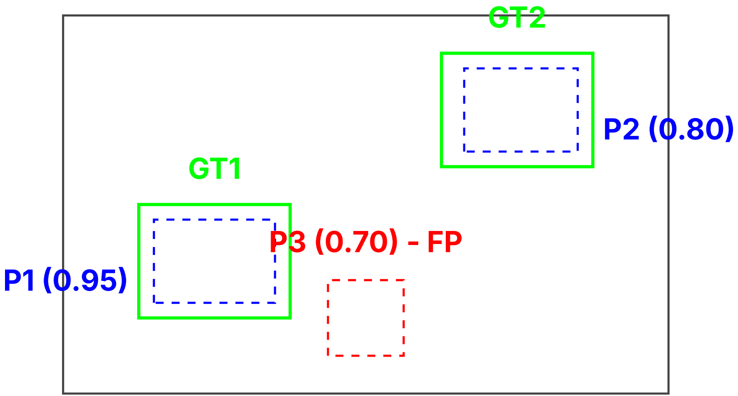
Analysis (IoU threshold = 0.5):

- P1 matches GT1: **TP**

Metrics:

- TP = 2, FP = 1, FN = 0

Example: Counting TP, FP, FN



Scenario: 2 GT objects, 3 predictions

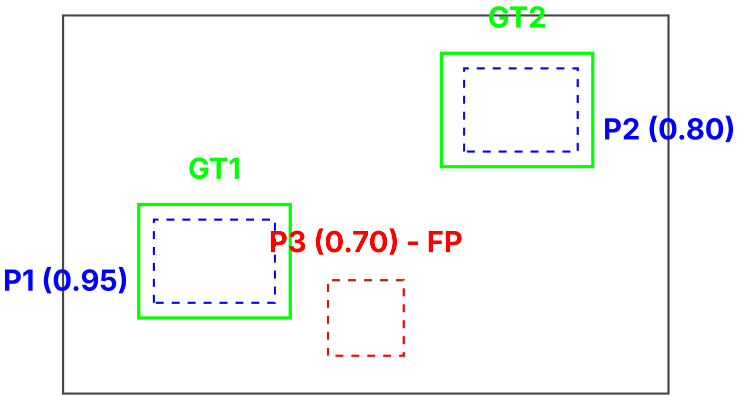
Analysis (IoU threshold = 0.5):

- P1 matches GT1: **TP**

Metrics:

- TP = 2, FP = 1, FN = 0

Example: Counting TP, FP, FN



Scenario: 2 GT objects, 3 predictions

Analysis (IoU threshold = 0.5):

- P1 matches GT1: **TP**

Metrics:

- TP = 2, FP = 1, FN = 0

Ranked Predictions Table

Example: Detection Results Sorted by Confidence

Given 5 predictions from our detector across the test set:

Confidence	Class	Box	TP/FP
0.95	Dog	(30,30,100,100)	TP
0.88	Bike	(150,120,200,180)	FP
0.80	Dog	(50,50,120,120)	TP
0.70	Person	(200,50,280,150)	TP
0.40	Cat	(100,100,150,150)	FP

Key Points

By varying the confidence threshold, we can trade off precision vs recall!

Precision-Recall Table

Threshold	Predictions	TP	FP	Precision	Recall
0.95	1	1	0	1.000	0.333
0.88	2	1	1	0.500	0.333
0.80	3	2	1	0.667	0.667
0.70	4	3	1	0.750	1.000
0.40	5	3	2	0.600	1.000

Assumptions: 3 ground truth objects total, IoU threshold = 0.5

- As threshold decreases → more predictions → recall increases

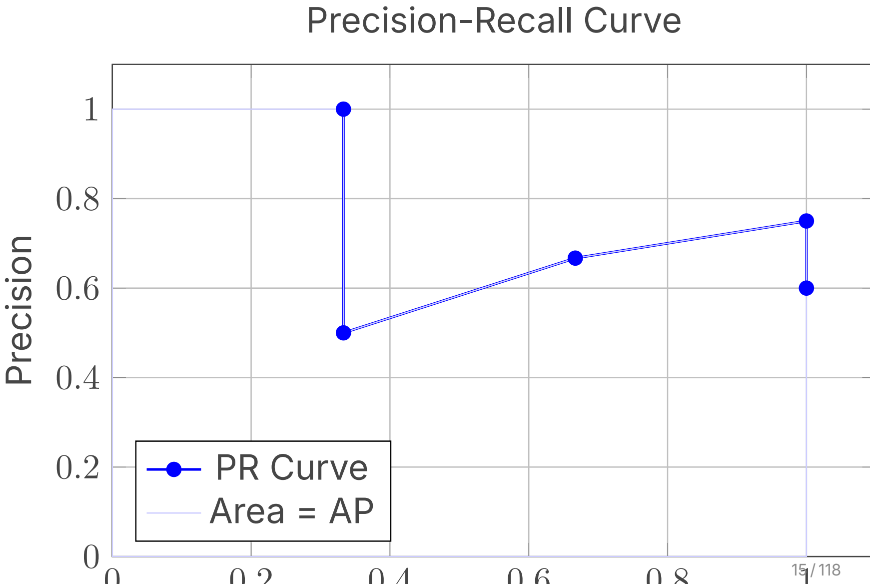
Precision-Recall Table

Threshold	Predictions	TP	FP	Precision	Recall
0.95	1	1	0	1.000	0.333
0.88	2	1	1	0.500	0.333
0.80	3	2	1	0.667	0.667
0.70	4	3	1	0.750	1.000
0.40	5	3	2	0.600	1.000

Assumptions: 3 ground truth objects total, IoU threshold = 0.5

- As threshold decreases → more predictions → recall increases
- But also more false positives → precision can decrease

Precision-Recall Curve



AP = Area under PR Curve

Definition: Average Precision Calculation

$$\text{AP} = \int_0^1 P(R)\, dR$$

In practice: Numerical integration or 11-point interpolation

AP = Area under PR Curve

Definition: Average Precision Calculation

$$\text{AP} = \int_0^1 P(R)\, dR$$

In practice: Numerical integration or 11-point interpolation

11-Point Interpolation:

AP = Area under PR Curve

Definition: Average Precision Calculation

$$\text{AP} = \int_0^1 P(R)\, dR$$

In practice: Numerical integration or 11-point interpolation

11-Point Interpolation:

- Sample at recall levels: 0, 0.1, 0.2, ..., 1.0

AP = Area under PR Curve

Definition: Average Precision Calculation

$$\text{AP} = \int_0^1 P(R)\, dR$$

In practice: Numerical integration or 11-point interpolation

11-Point Interpolation:

- Sample at recall levels: 0, 0.1, 0.2, ..., 1.0
- For each recall r, find max precision for recall $\geq r$

AP = Area under PR Curve

Definition: Average Precision Calculation

$$\text{AP} = \int_0^1 P(R)\, dR$$

In practice: Numerical integration or 11-point interpolation

11-Point Interpolation:

- Sample at recall levels: 0, 0.1, 0.2, ..., 1.0
- For each recall r, find max precision for recall $\geq r$
- Average the 11 precision values

AP = Area under PR Curve

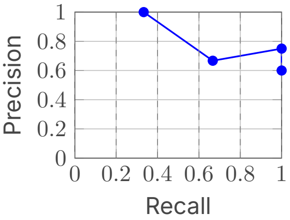
Definition: Average Precision Calculation

$$\text{AP} = \int_0^1 P(R)\, dR$$

In practice: Numerical integration or 11-point interpolation

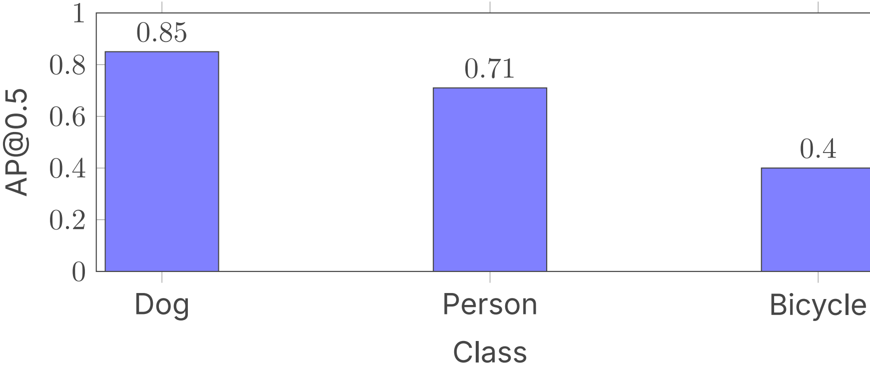
11-Point Interpolation:

- Sample at recall levels: 0, 0.1, 0.2, ..., 1.0
- For each recall r, find max precision for recall $\geq r$
- Average the 11 precision values



Class-wise AP Example

Class	AP@0.5	Visual
Dog	0.85	Excellent
Person	0.71	Good
Bicycle	0.40	Poor



Interpretation: Dog detection works well, but bicycle

Mean Average Precision (mAP)

Definition: mAP Calculation

$$\text{mAP} = \frac{1}{C}\sum_{c=1}^{C} \text{AP}_c$$

where C is the number of classes

Example: Our Example

$$\text{mAP} = \frac{\text{AP}_{dog} + \text{AP}_{person} + \text{AP}_{bicycle}}{3}$$

$$= \frac{0.85 + 0.71 + 0.40}{3} = \mathbf{0.653}$$

Class-Agnostic mAP (CA-mAP)

Definition: Class-Agnostic Evaluation

Ignore class labels when matching predictions to ground truth.
Match based on IoU overlap alone.

Class-Agnostic mAP (CA-mAP)

Definition: Class-Agnostic Evaluation

Ignore class labels when matching predictions to ground truth.
Match based on IoU overlap alone.

Standard mAP:

Class-Agnostic mAP:

Class-Agnostic mAP (CA-mAP)

Definition: Class-Agnostic Evaluation

Ignore class labels when matching predictions to ground truth.
Match based on IoU overlap alone.

Standard mAP:

- Dog pred $\leftrightarrow$ Dog GT: ✓

Class-Agnostic mAP:

Class-Agnostic mAP (CA-mAP)

Definition: Class-Agnostic Evaluation

Ignore class labels when matching predictions to ground truth.
Match based on IoU overlap alone.

Standard mAP:

- Dog pred $\leftrightarrow$ Dog GT: ✓
- Dog pred $\leftrightarrow$ Person GT: $\times$

Class-Agnostic mAP:

Class-Agnostic mAP (CA-mAP)

Definition: Class-Agnostic Evaluation

Ignore class labels when matching predictions to ground truth.
Match based on IoU overlap alone.

Standard mAP:

- Dog pred $\leftrightarrow$ Dog GT: $\checkmark$
- Dog pred $\leftrightarrow$ Person GT: $\times$

Class-Agnostic mAP:

- Any pred $\leftrightarrow$ Any GT (if IoU > threshold): $\checkmark$

Class-Agnostic mAP (CA-mAP)

Definition: Class-Agnostic Evaluation

Ignore class labels when matching predictions to ground truth.
Match based on IoU overlap alone.

Standard mAP:

- Dog pred $\leftrightarrow$ Dog GT: ✓
- Dog pred $\leftrightarrow$ Person GT: $\times$

Class-Agnostic mAP:

- Any pred $\leftrightarrow$ Any GT (if IoU > threshold): ✓
- Useful for generic object

Class-Agnostic mAP (CA-mAP)

Definition: Class-Agnostic Evaluation

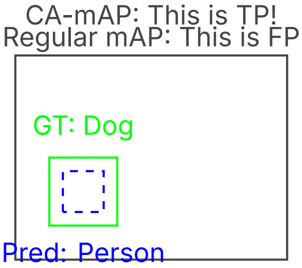
Ignore class labels when matching predictions to ground truth.
Match based on IoU overlap alone.

Standard mAP:

- Dog pred ↔ Dog GT: ✓
- Dog pred ↔ Person GT: ×

Class-Agnostic mAP:

- Any pred ↔ Any GT (if IoU > threshold): ✓
- Useful for generic object



Strict Evaluation: COCO Metrics

Definition: COCO Evaluation Protocol

- **AP@50**: IoU threshold = 0.5 (lenient)

Metric	Value	Interpretation
mAP@50	0.71	Good localization (loose)
mAP@75	0.45	Moderate precise localization
mAP@[.5:.95]	0.42	Overall localization quality

Key Points

Strict Evaluation: COCO Metrics

Definition: COCO Evaluation Protocol

- **AP@50**: IoU threshold = 0.5 (lenient)
- **AP@75**: IoU threshold = 0.75 (strict)

Metric	Value	Interpretation
mAP@50	0.71	Good localization (loose)
mAP@75	0.45	Moderate precise localization
mAP@[.5:.95]	0.42	Overall localization quality

Key Points

Key Insight: Higher IoU thresholds demand more pre

Strict Evaluation: COCO Metrics

Definition: COCO Evaluation Protocol

- **AP@50**: IoU threshold = 0.5 (lenient)
- **AP@75**: IoU threshold = 0.75 (strict)
- **AP@[.5:.95]**: Average over IoU thresholds 0.5, 0.55, 0.6, ..., 0.95

Metric	Value	Interpretation
mAP@50	0.71	Good localization (loose)
mAP@75	0.45	Moderate precise localization
mAP@[.5:.95]	0.42	Overall localization quality

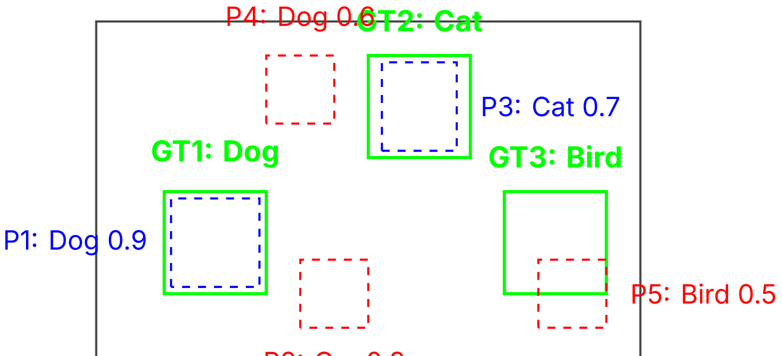
Key Points

Key Insight: Higher IoU thresholds demand more pre

Pop Quiz 1: Compute Precision & Recall

Quick Quiz Pop Quiz 0

Given the detection scenario below, compute precision and recall (IoU threshold = 0.5):



Pop Quiz 1: Answer

Example: Solution

Analysis (with IoU > 0.5 matching):

- P1 (Dog 0.9) matches GT1 (Dog): **TP**

Final counts: TP = 2, FP = 3, FN = 1

$$\text{Precision} = \frac{2}{2+3} = 0.40 \quad \text{Recall} = \frac{2}{2+1} = 0.67$$

Pop Quiz 1: Answer

Example: Solution

Analysis (with IoU > 0.5 matching):

- P1 (Dog 0.9) matches GT1 (Dog): **TP**
- P2 (Car 0.8) no GT match: **FP**

Final counts: TP = 2, FP = 3, FN = 1

$$\text{Precision} = \frac{2}{2+3} = 0.40 \quad \text{Recall} = \frac{2}{2+1} = 0.67$$

Pop Quiz 1: Answer

Example: Solution

Analysis (with IoU > 0.5 matching):

- P1 (Dog 0.9) matches GT1 (Dog): **TP**
- P2 (Car 0.8) no GT match: **FP**
- P3 (Cat 0.7) matches GT2 (Cat): **TP**

Final counts: TP = 2, FP = 3, FN = 1

$$\text{Precision} = \frac{2}{2+3} = 0.40 \quad \text{Recall} = \frac{2}{2+1} = 0.67$$

Pop Quiz 1: Answer

Example: Solution

Analysis (with IoU > 0.5 matching):

- P1 (Dog 0.9) matches GT1 (Dog): **TP**
- P2 (Car 0.8) no GT match: **FP**
- P3 (Cat 0.7) matches GT2 (Cat): **TP**
- P4 (Dog 0.6) no GT match: **FP**

Final counts: TP = 2, FP = 3, FN = 1

$$\text{Precision} = \frac{2}{2+3} = 0.40 \quad \text{Recall} = \frac{2}{2+1} = 0.67$$

Pop Quiz 1: Answer

Example: Solution

Analysis (with IoU > 0.5 matching):

- P1 (Dog 0.9) matches GT1 (Dog): **TP**
- P2 (Car 0.8) no GT match: **FP**
- P3 (Cat 0.7) matches GT2 (Cat): **TP**
- P4 (Dog 0.6) no GT match: **FP**
- P5 (Bird 0.5) poor overlap with GT3: **FP**

Final counts: TP = 2, FP = 3, FN = 1

$$\text{Precision} = \frac{2}{2+3} = 0.40 \quad \text{Recall} = \frac{2}{2+1} = 0.67$$

Pop Quiz 1: Answer

Example: Solution

Analysis (with IoU > 0.5 matching):

- P1 (Dog 0.9) matches GT1 (Dog): **TP**
- P2 (Car 0.8) no GT match: **FP**
- P3 (Cat 0.7) matches GT2 (Cat): **TP**
- P4 (Dog 0.6) no GT match: **FP**
- P5 (Bird 0.5) poor overlap with GT3: **FP**
- GT3 (Bird) unmatched: **FN**

Final counts: TP = 2, FP = 3, FN = 1

$$\text{Precision} = \frac{2}{2+3} = 0.40 \quad \text{Recall} = \frac{2}{2+1} = 0.67$$

Summary Table

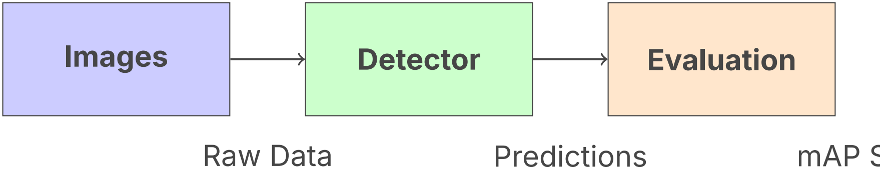
Concept	Meaning	Key Insight
IoU	Box overlap quality	Matching criterion (usua
Precision	Detection quality	$\frac{TP}{TP+FP}$ (fewer false alarm
Recall	Detection coverage	$\frac{TP}{TP+FN}$ (fewer missed ob
AP	Area under PR curve	Single-class performanc
mAP	Average AP over classes	Multi-class detector per
CA-mAP	Class-agnostic mAP	Localization-only evalua
COCO	Multi-IoU evaluation	AP@[.5:.95] for precise

Key Points

Golden Rule "Detection is not just about finding objects, but finding them right."

What We've Learned

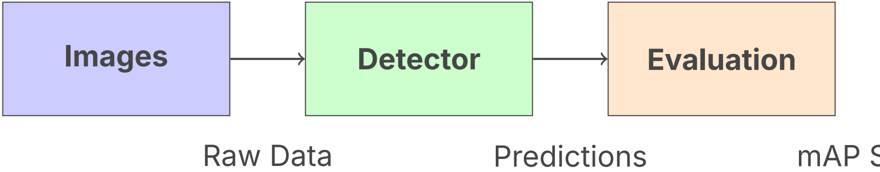
- **Task hierarchy**: Classification → Localization → Detection



Next steps: Explore modern architectures (YOLO,

What We've Learned

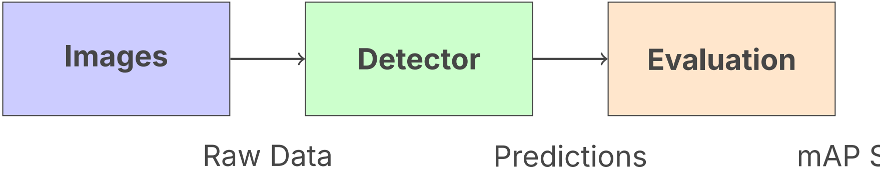
- **Task hierarchy:** Classification → Localization → Detection
- **Evaluation pipeline:** IoU matching → TP/FP counting → PR curves → AP/mAP



Next steps: Explore modern architectures (YOLO,

What We've Learned

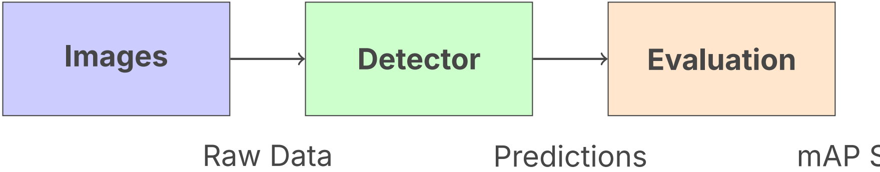
- **Task hierarchy**: Classification → Localization → Detection
- **Evaluation pipeline**: IoU matching → TP/FP counting → PR curves → AP/mAP
- **Trade-offs**: Precision vs Recall, lenient vs strict IoU thresholds



Next steps: Explore modern architectures (YOLO,

What We've Learned

- **Task hierarchy**: Classification → Localization → Detection
- **Evaluation pipeline**: IoU matching → TP/FP counting → PR curves → AP/mAP
- **Trade-offs**: Precision vs Recall, lenient vs strict IoU thresholds
- **Practical metrics**: COCO-style evaluation for real-world deployment



Next steps: Explore modern architectures (YOLO,